



**UNIVERSITÄT
HEIDELBERG**
ZUKUNFT
SEIT 1386

An introduction to Large Language Models and their biases

15.01.2025

Presentation for the course “El español en el mundo: variación, traducción y nuevas tecnologías”

Prof. Dr. Vanesa Rodríguez Tembrás

BA Übersetzungswissenschaft - Heidelberg University

Guest lecturer:

Mauricio Fadel Argerich

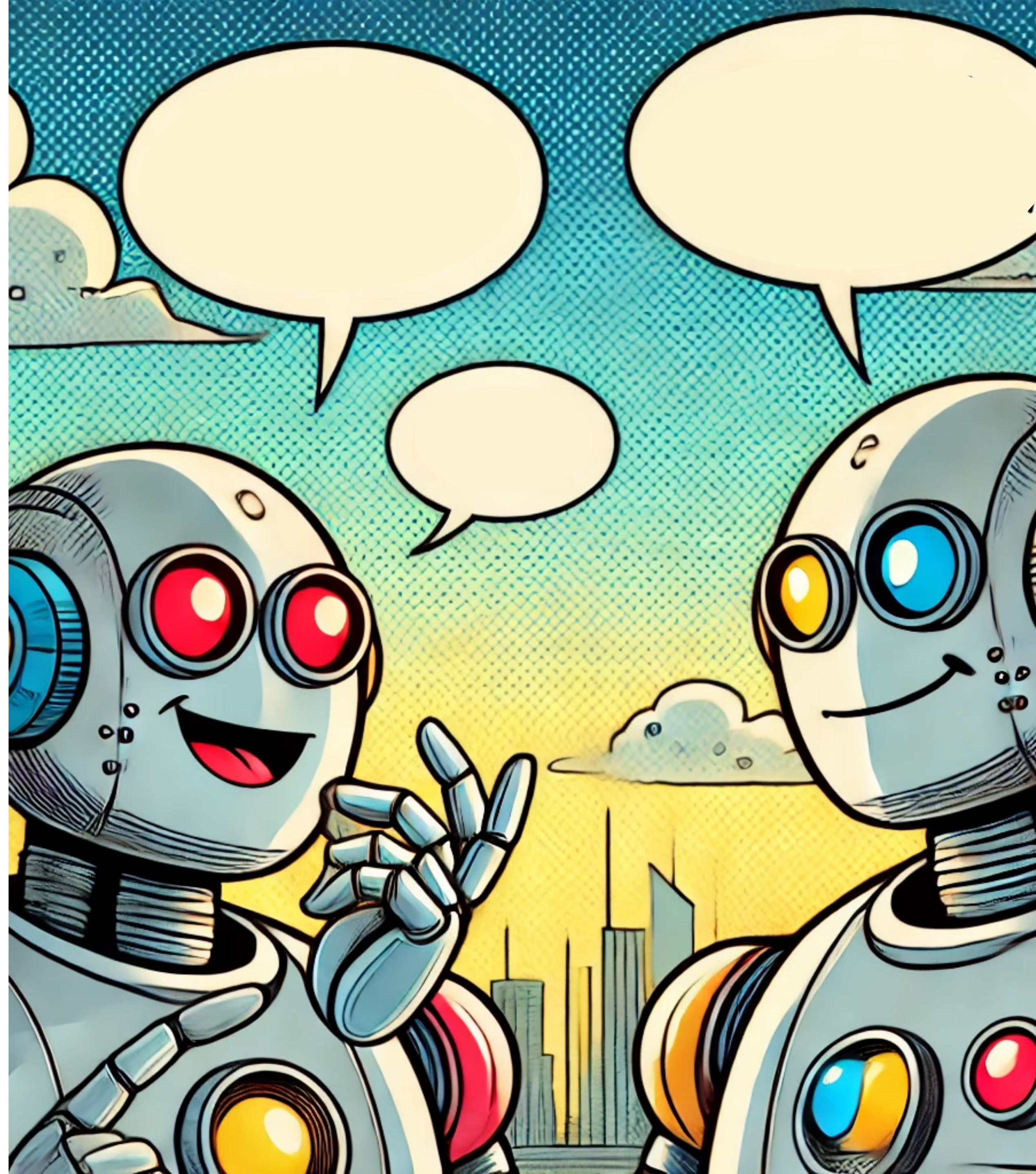
Tech Strategy & Innovation Manager @ O2 Telefonica Germany

PhD Candidate @ Universidad Politécnica de Madrid

mauricio.fadel@alumnos.upm.es

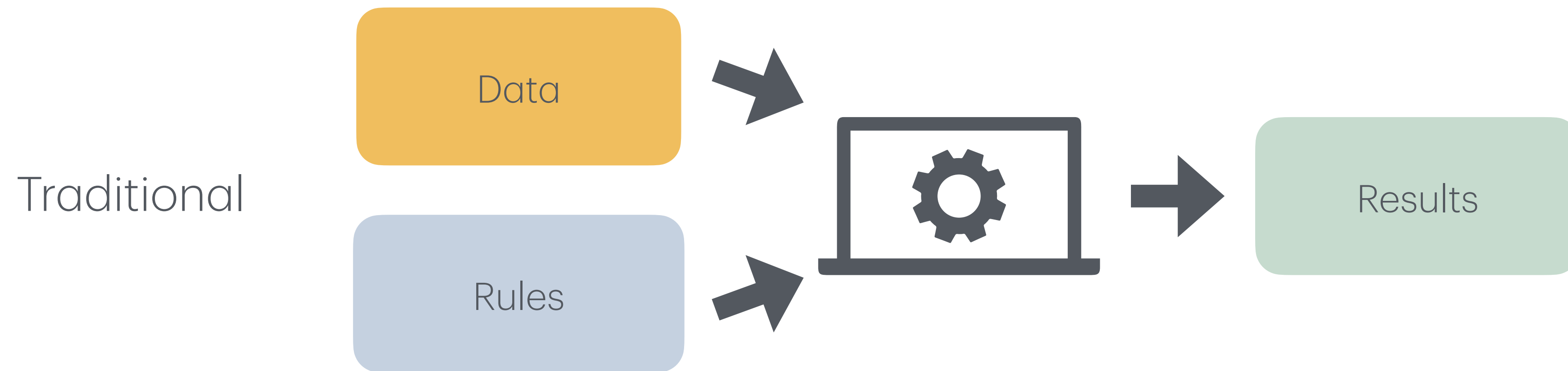
Agenda

- Introduction to ML
- Errors in ML: bias, variance, data bias
- LLMs
 - Introduction to LLMs
 - The task
- Biases of LLMs
- Reducing bias in LLMs
- Wrap up



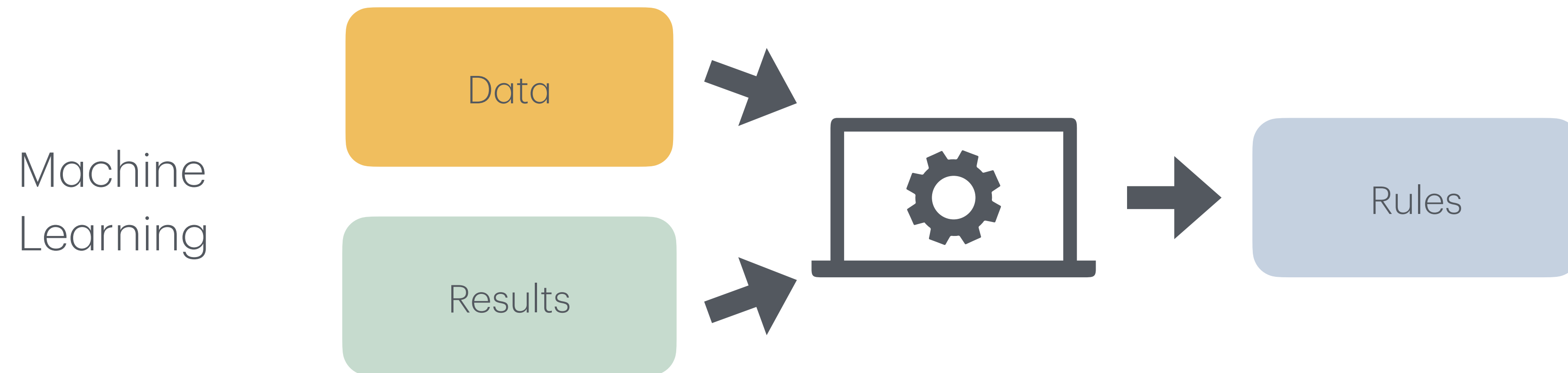
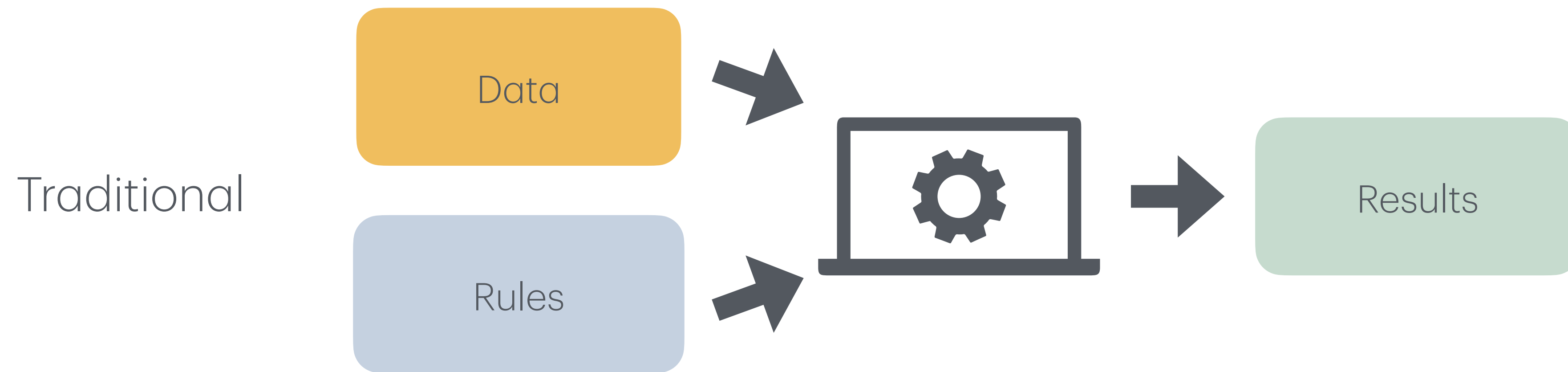
Introduction to Machine Learning

What's different from traditional programming?



Introduction to Machine Learning

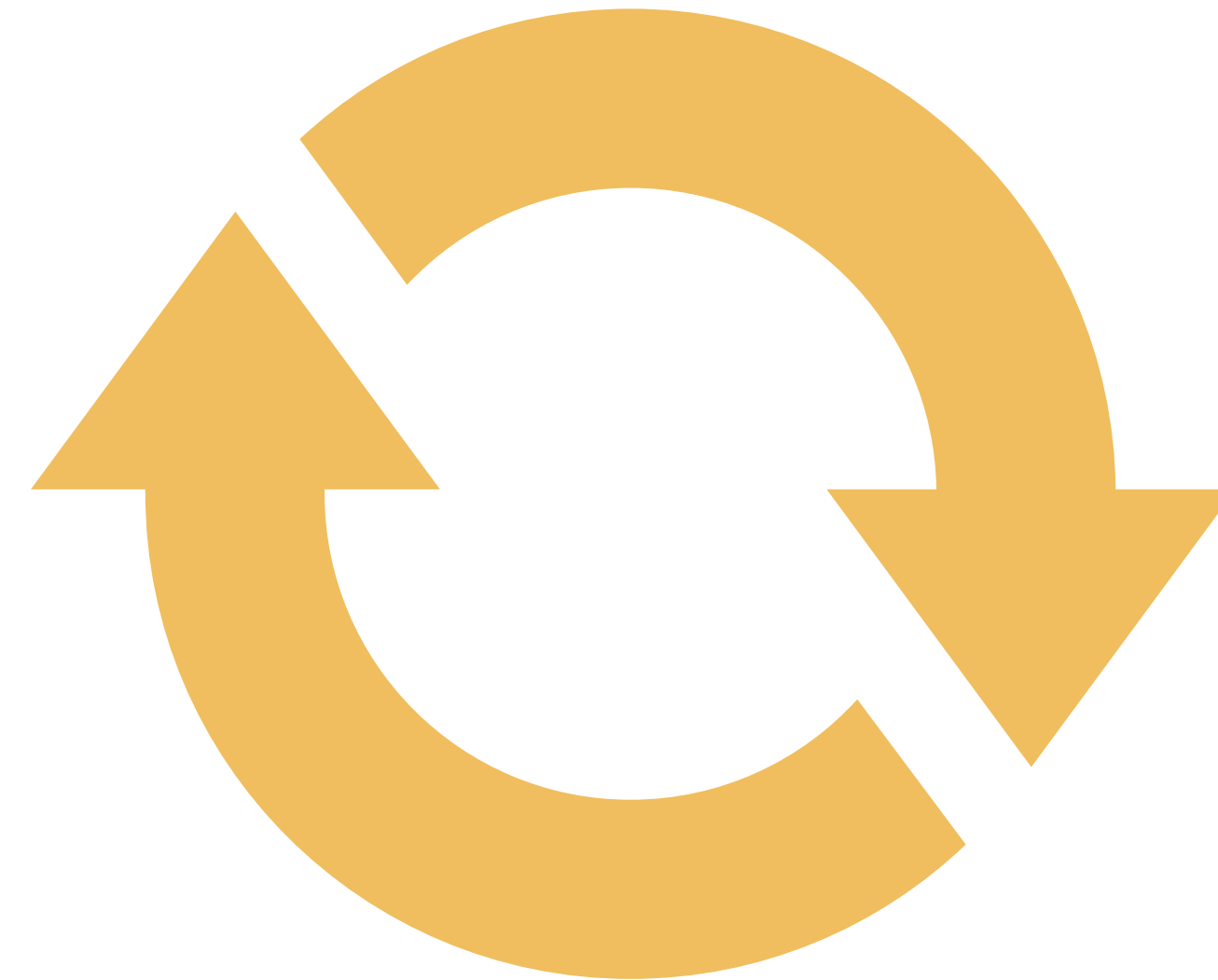
What's different from traditional programming?



Introduction to Machine Learning

Lifecycle of a ML model

1. Problem definition
2. Data collection and preprocessing
3. Training
4. Testing/Evaluation
5. Inference
6. Monitoring and updating



Introduction to Machine Learning

How does the machine learn the rules?

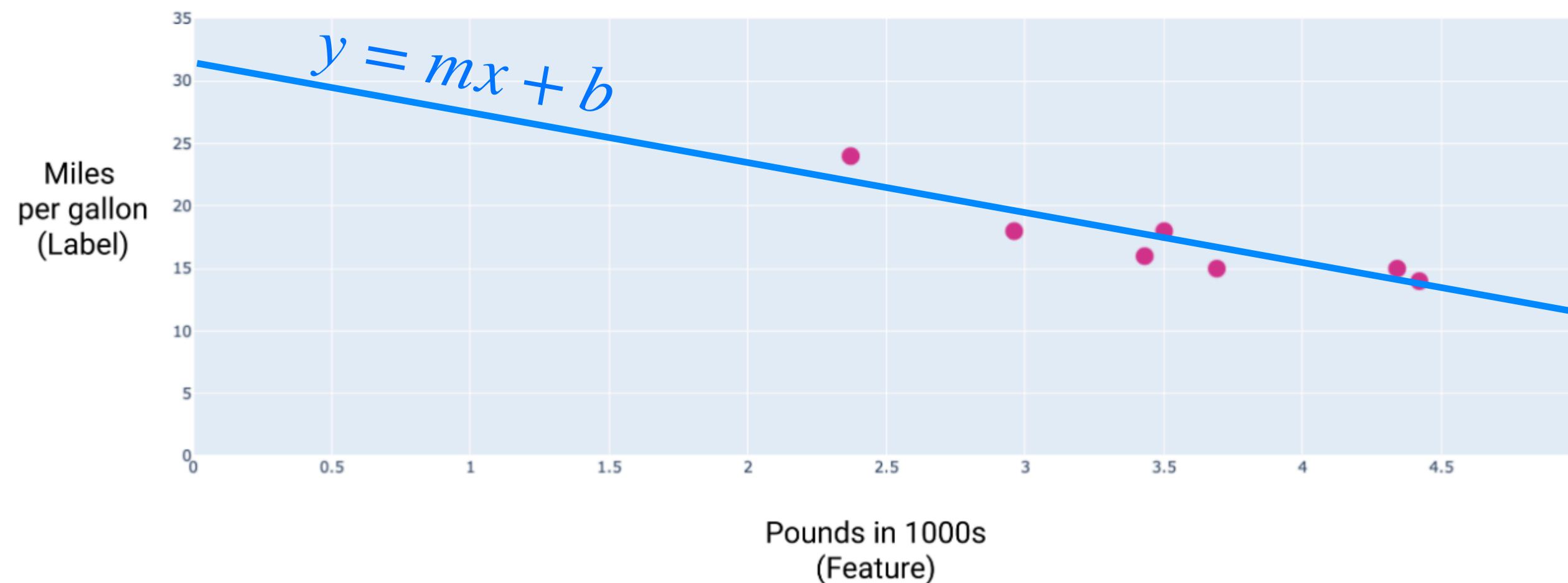
- Let's see how a simple approach works: **Linear Regression (LR)**
- In ML, LR finds the relationship between features and a label
- E.g., we want to predict a car's fuel efficiency in miles per gallon based on how heavy the car is. We have the following dataset:

Pounds in 1000s (feature)	Miles per gallon (label)
3.5	18
3.69	15
3.44	18
3.43	16
4.34	15
4.42	14
2.37	24

Introduction to Machine Learning

How does the machine learn the rules?

- If we plot these data, we see there's a relationship between the features and the labels
- We can create a predicting model by drawing a best fit line through the points



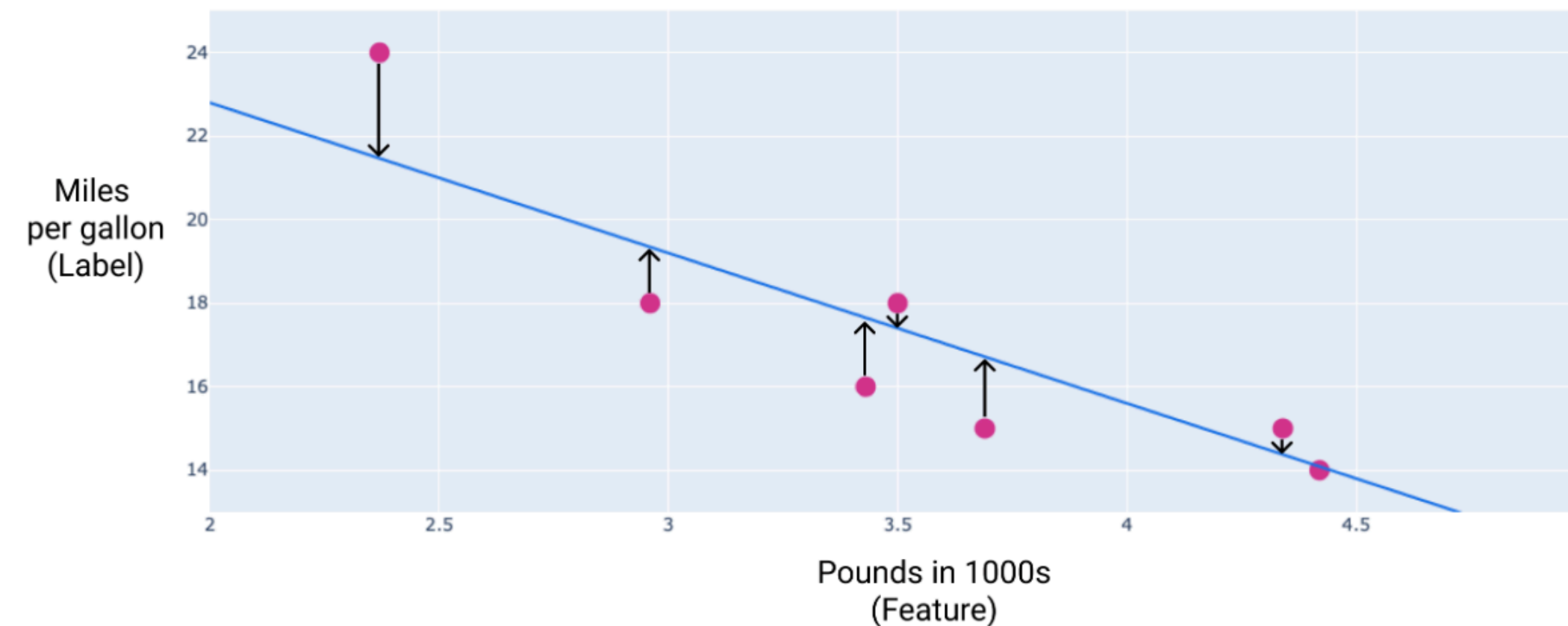
y is miles per gallon—the value we want to predict.
 m is the slope of the line.
 x is pounds—our input value.
 b is the y-intercept.

- During training, the model calculates the **weight** and **bias** to produce the best model

Introduction to Machine Learning

How does the machine learn the rules?

- To calculate the best **weight** and **bias**, the machine uses the **loss function**
- The **loss** defines how wrong the model's predictions are

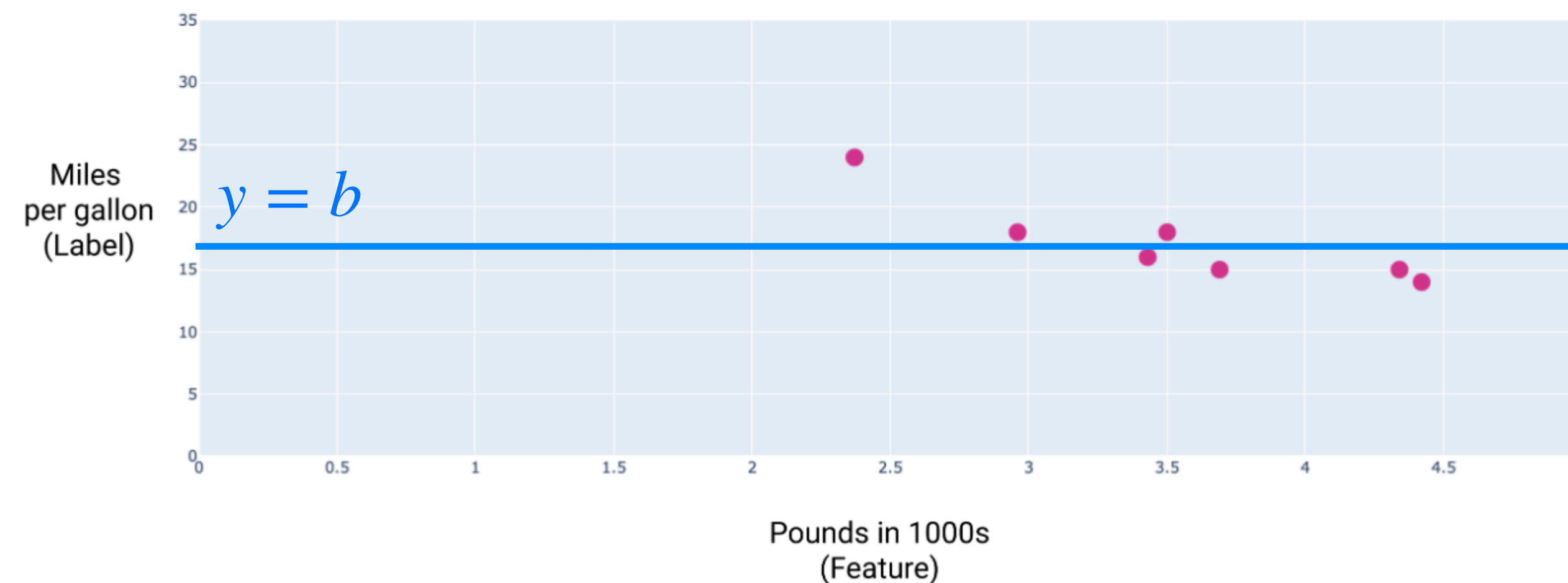


- When learning, the model starts with random values for the weight and bias, and using gradient descent, iteratively updates their values to find the best fit line (model)

Errors in ML: bias, variance, data bias

Looking more deeply into prediction errors

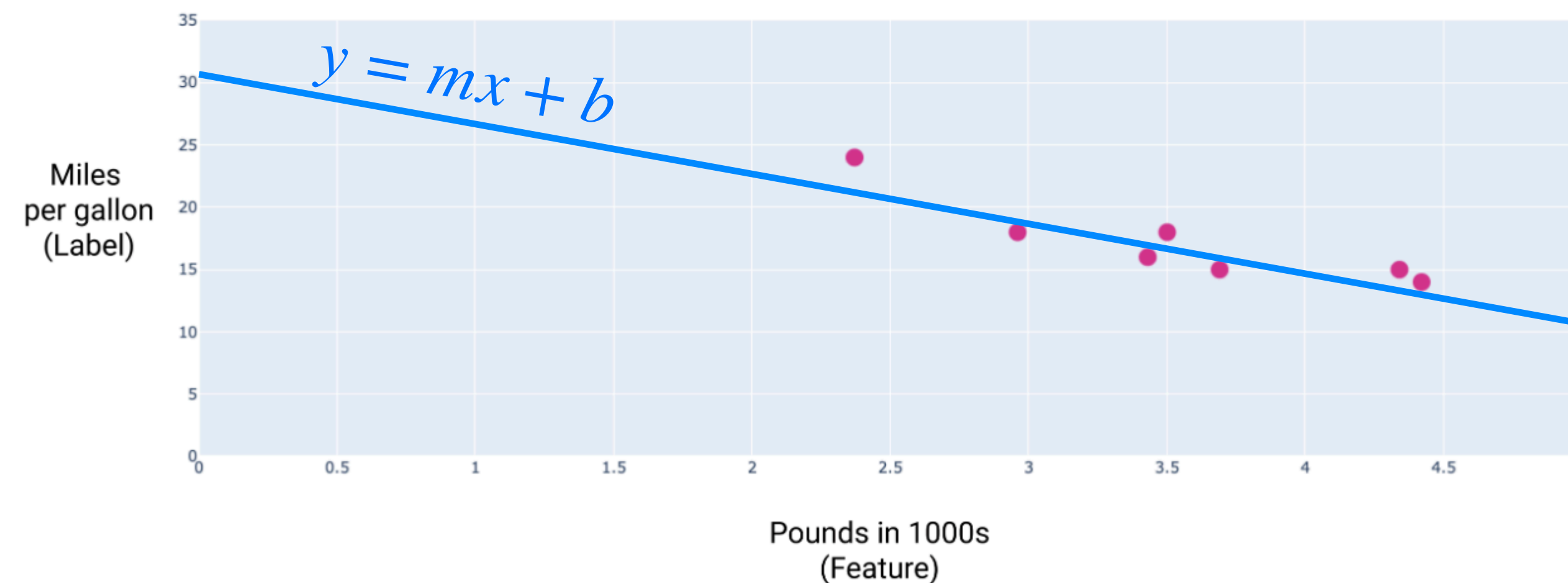
- As you've seen, a model is nothing else than a mathematical function: we can use many different models as well as features to improve our predictions



Errors in ML: bias, variance, data bias

Looking more deeply into prediction errors

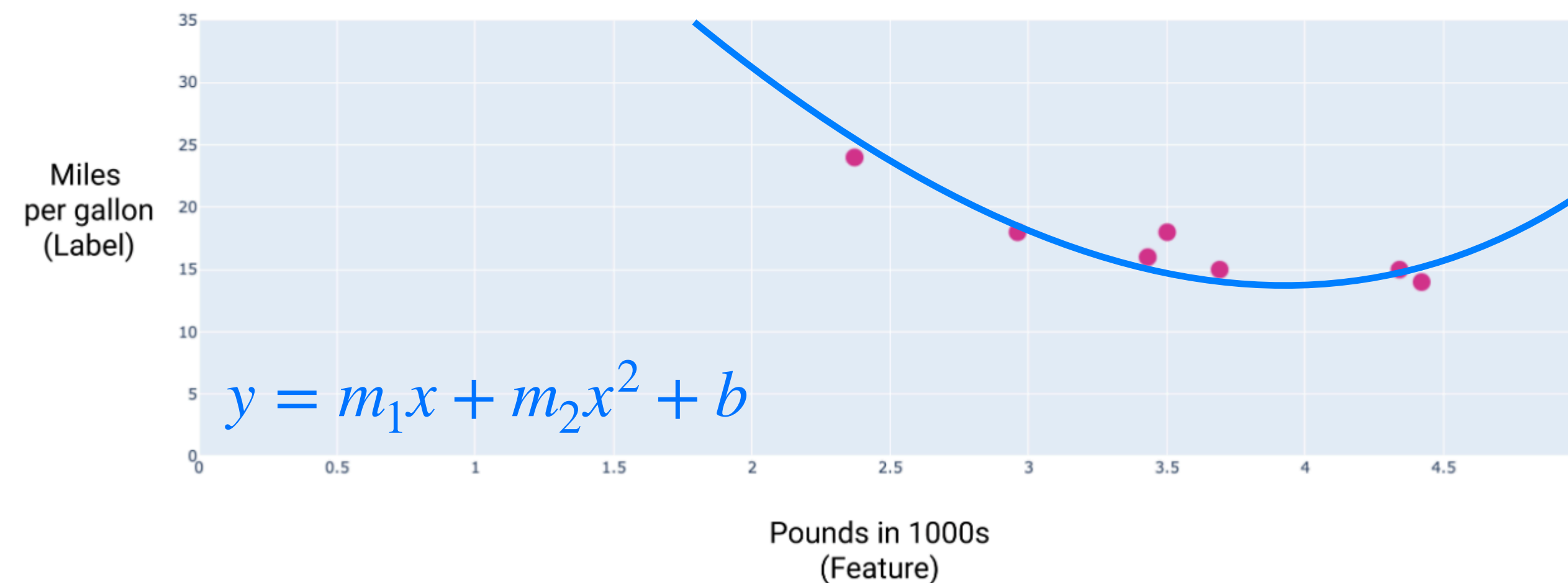
- As you've seen, a model is nothing else than a mathematical function: we can use many different models as well as features to improve our predictions



Errors in ML: bias, variance, data bias

Looking more deeply into prediction errors

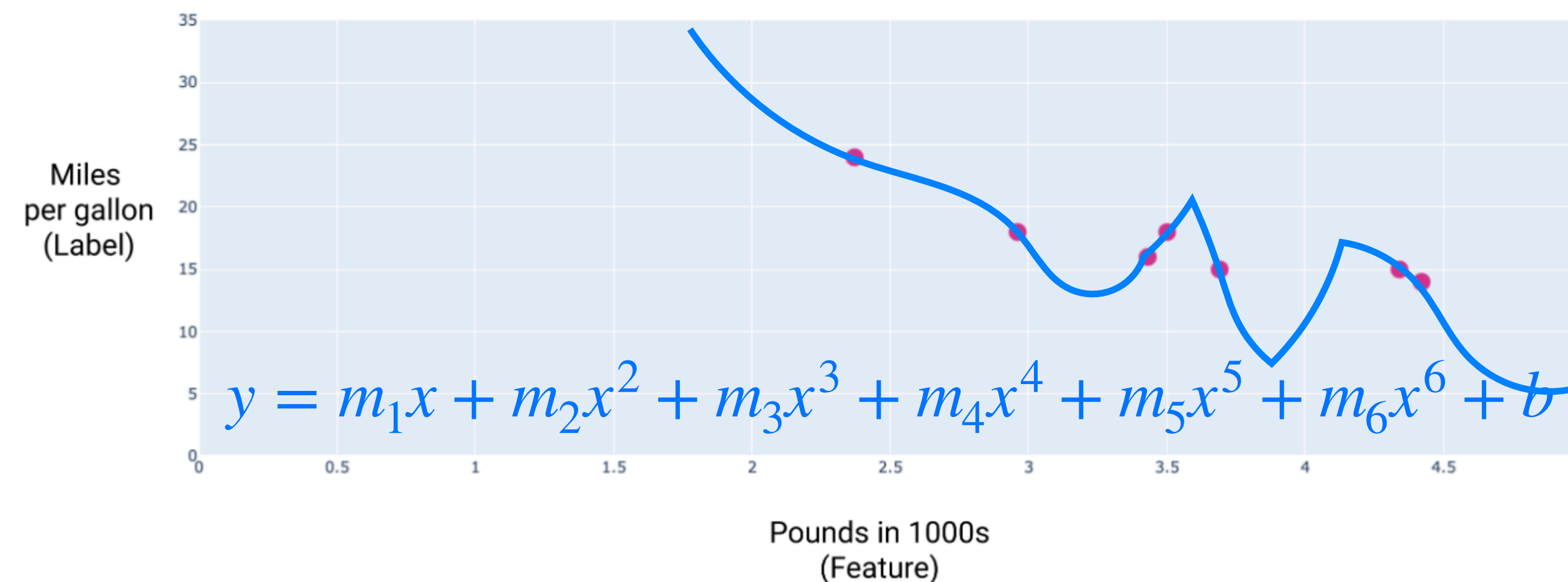
- As you've seen, a model is nothing else than a mathematical function: we can use many different models as well as features to improve our predictions



Errors in ML: bias, variance, data bias

Looking more deeply into prediction errors

- As you've seen, a model is nothing else than a mathematical function: we can use many different models as well as features to improve our predictions



Errors in ML: bias, variance, data bias

Looking more deeply into prediction errors

- As you've seen, a model is nothing else than a mathematical function: we can use many different models as well as features to improve our predictions

$$y' = b + w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4 + w_5 X_5$$

The diagram shows a linear regression equation with five terms. Below the equation, five labels are connected to their respective terms by lines: 'Pounds' connects to $w_1 X_1$, 'Displacement' connects to $w_2 X_2$, 'Acceleration' connects to $w_3 X_3$, 'Number of cylinders' connects to $w_4 X_4$, and 'Horsepower' connects to $w_5 X_5$.

Errors in ML: bias, variance, data bias

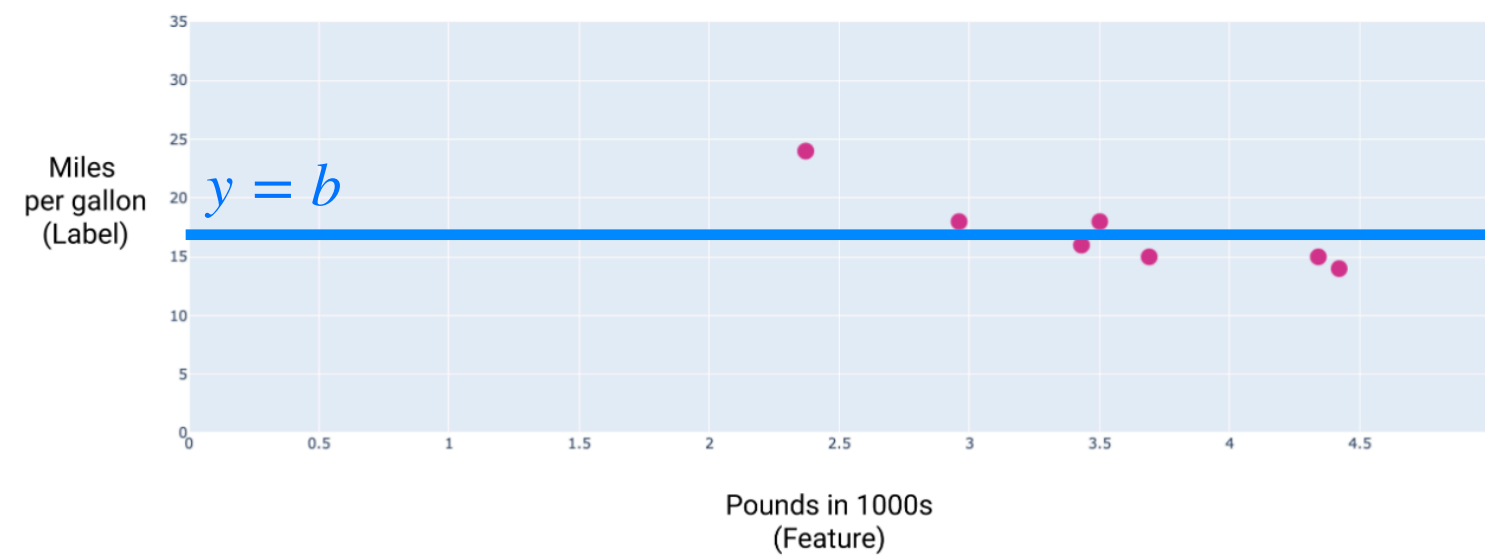
The bias-variance tradeoff

- The error comes from two factors: **bias** and **variance**
- **Bias**
 - The systematic error due to overly simple **assumptions in the model**.
 - Leads to **underfitting**: the model fails to capture the underlying patterns in the data.
- **Variance**
 - Error due to **sensitivity** to fluctuations in **the training data**.
 - Leads to **overfitting**: the model captures noise rather than the true pattern

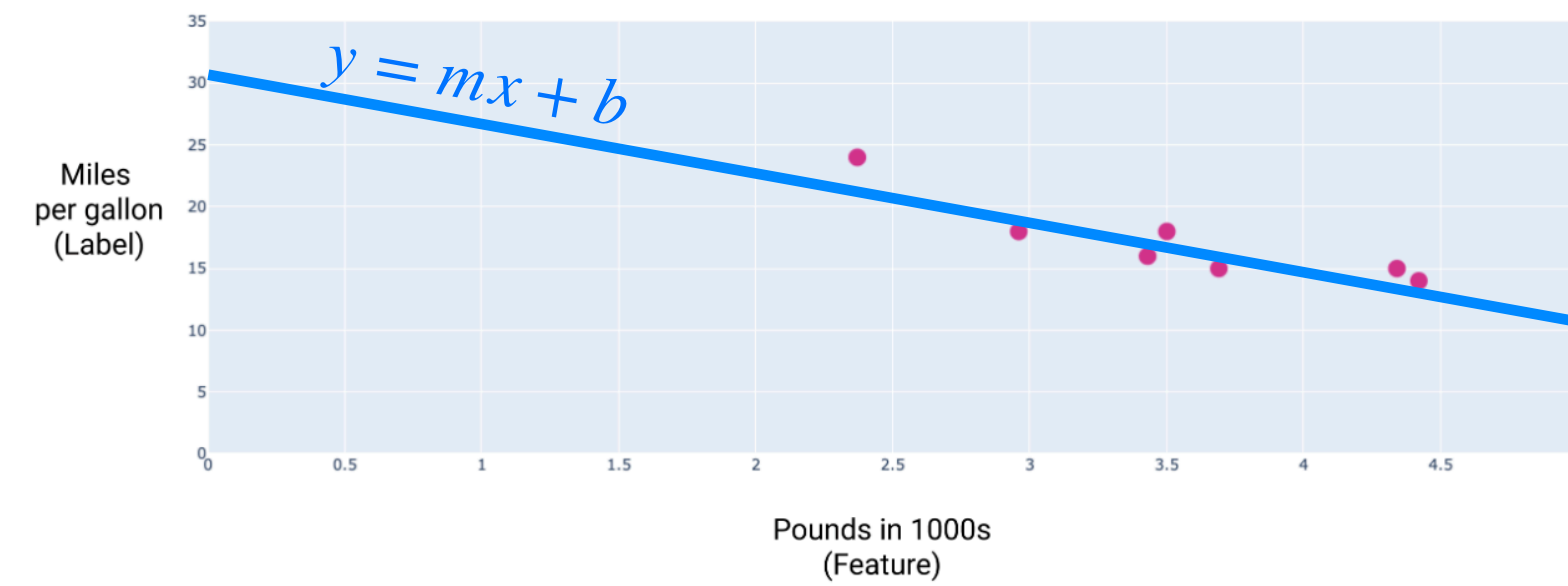
Errors in ML: bias, variance, data bias

The bias-variance tradeoff

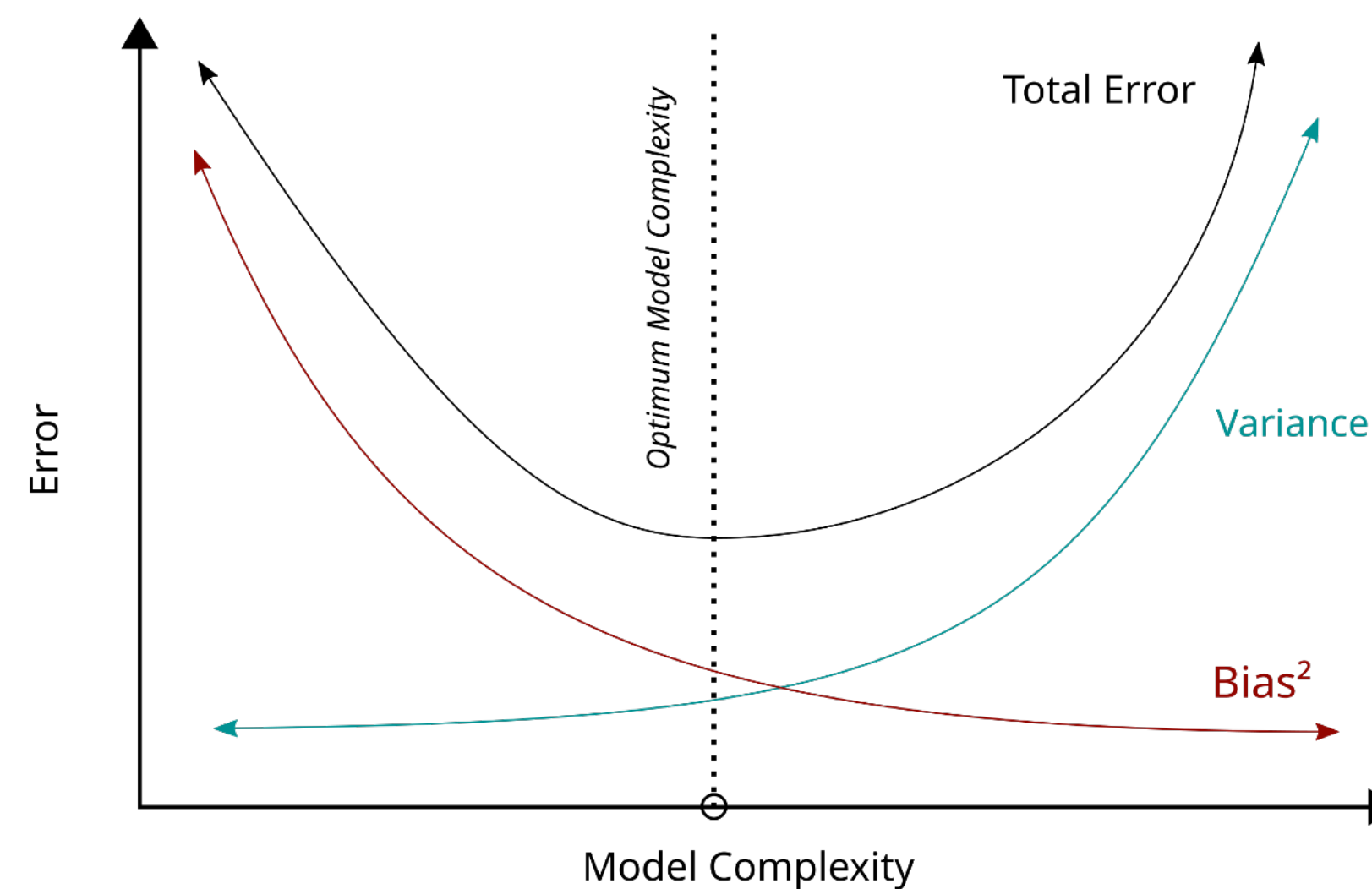
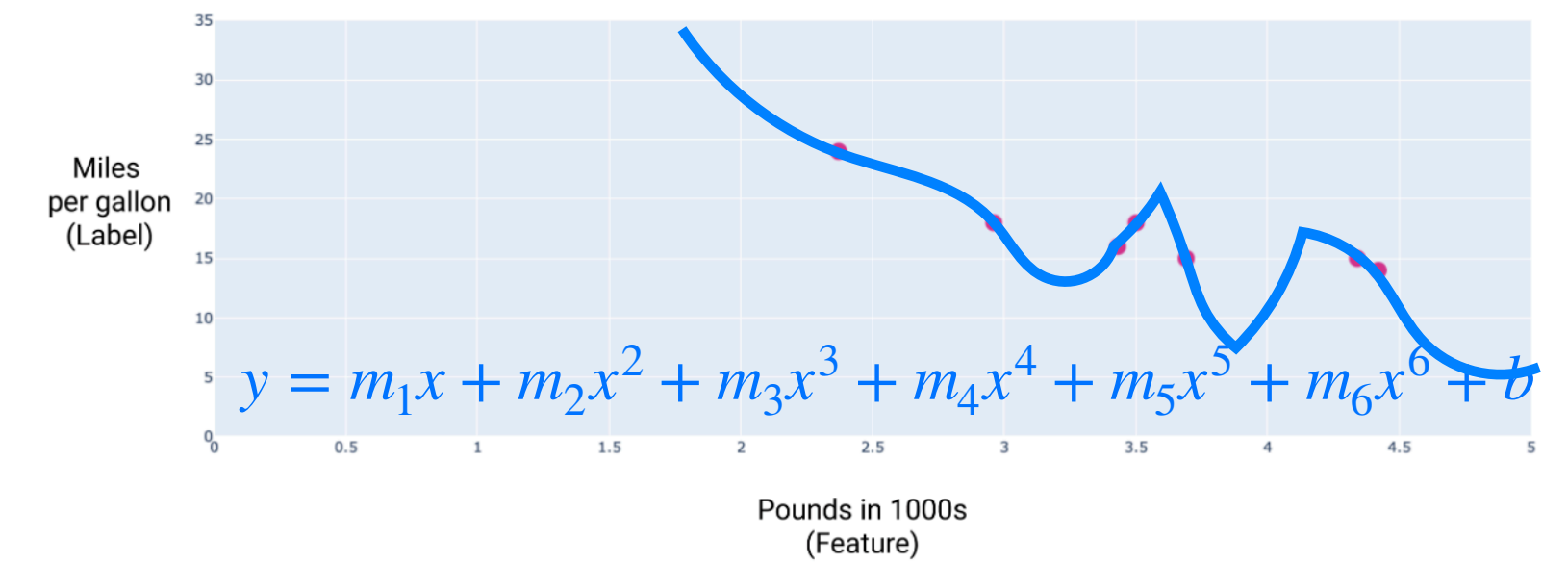
Underfitting



Just right



Overfitting



Errors in ML: bias, variance, data bias

Data bias

- We just talked about the errors coming from assumptions in the model, but there are also errors that come from assumptions in the (training) data
- From previous example: data from only one car, data from one car manufacturer, etc.
- The model only learns what it “sees” in the training data

Types of data bias:

Confirmation bias

Historical bias

Selection/sampling bias

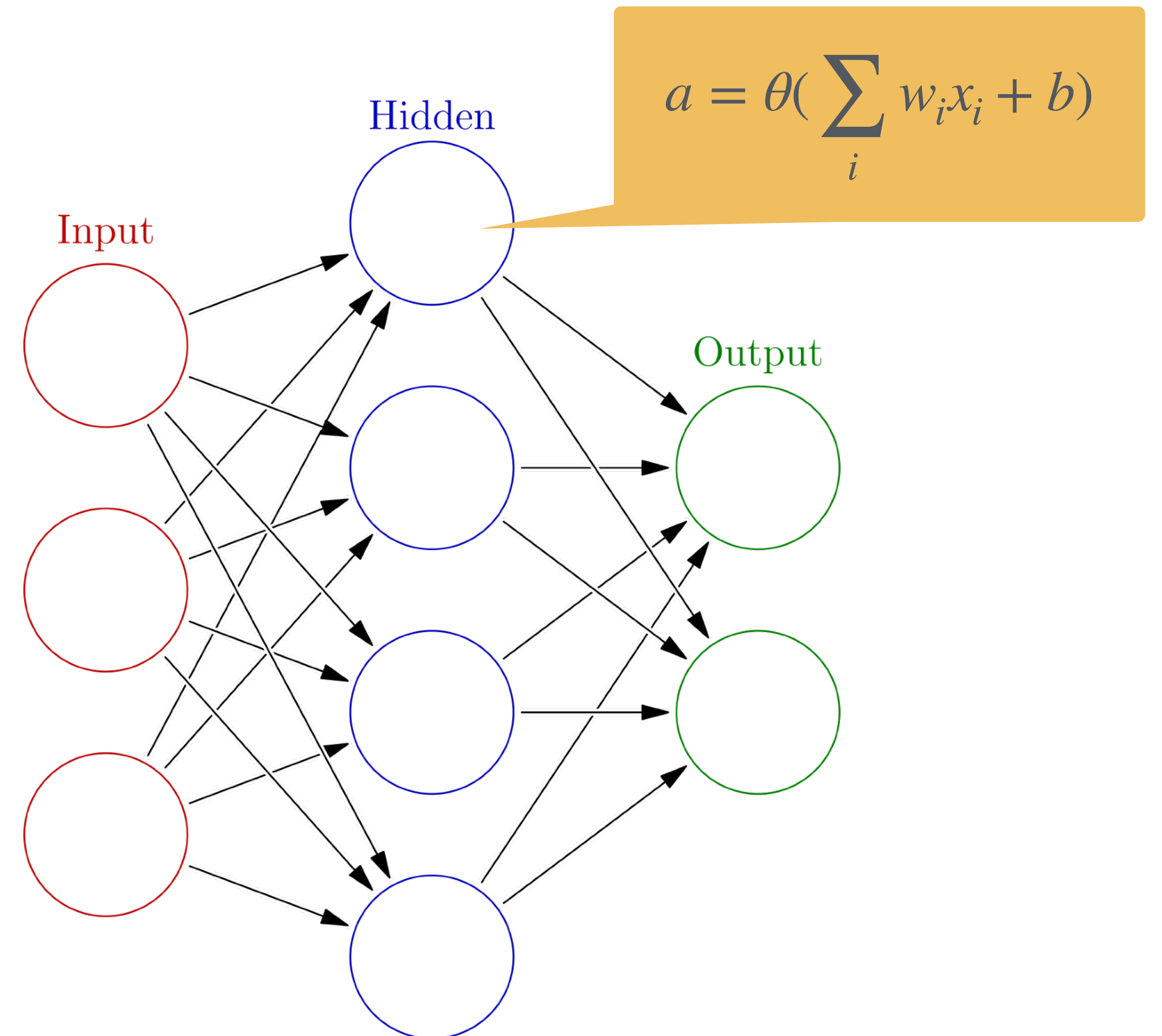
Survivorship bias

Availability bias

LLMs (Large Language Models): Introduction

The neural network

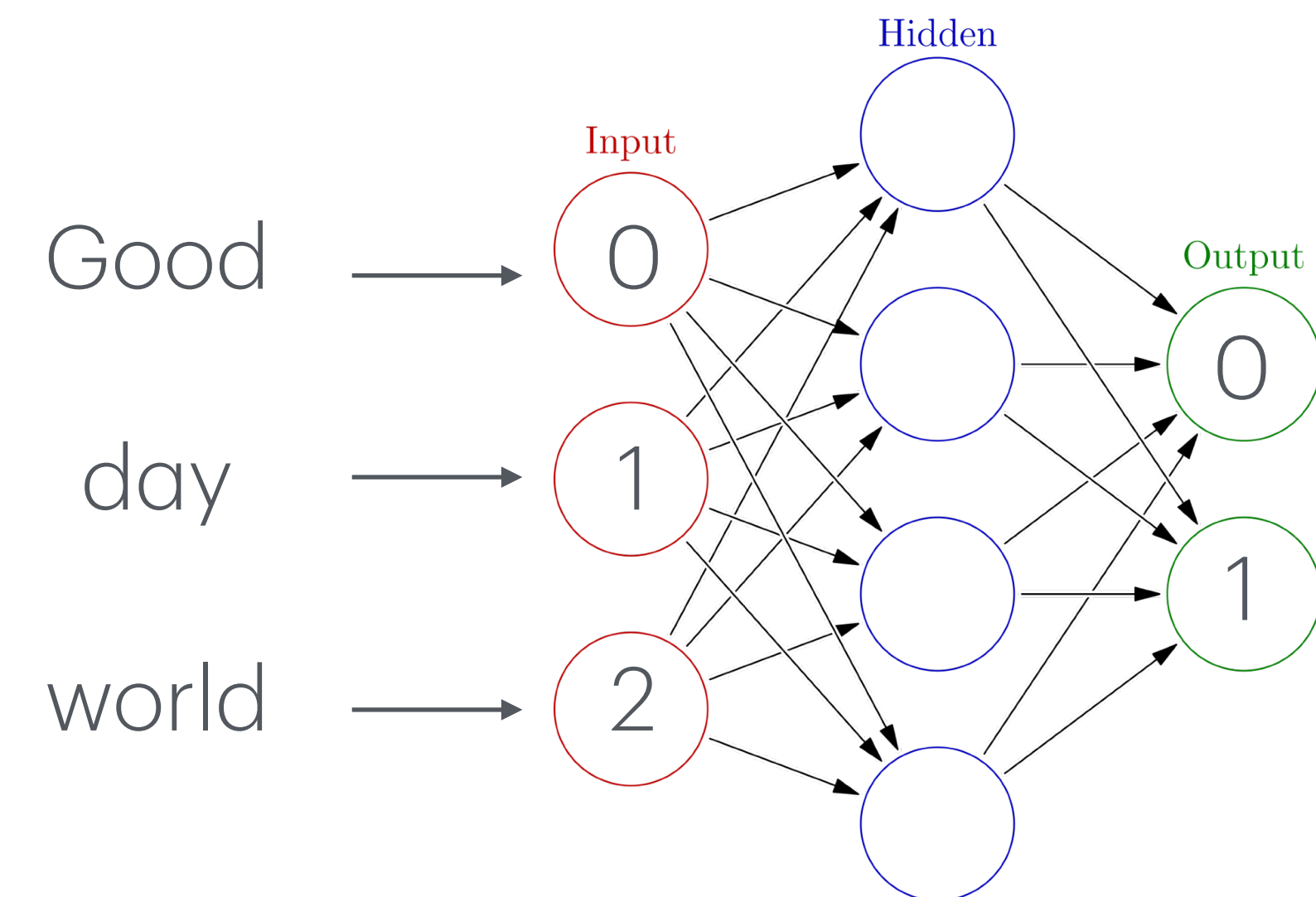
- **Neural networks (NN)** are a type of ML model
- Roughly, each neuron can be thought of a linear regression, to which a **non-linear function** is applied
- During training, the weights for all the neurons are learned, these are the **parameters** of the NN
- Therefore, a NN is a large mathematical function



LLMs (Large Language Models): Introduction

Making NNs talk

- NNs can process **text** if we convert each word (or token) into numbers
- The output is also composed of numbers that we can then convert back to text
- There are several **different NN architectures** and some of them have shown excellent results on natural language
- In general, the larger the NN, the better results it can achieve
 - > LLMs are really **large** with GPT-4 estimated to have 1.8 trillion parameters!



LLMs: The Task

Even though LLMs are powerful, they are trained for a very simple task: predicting the next word

The images in the presentation were fun
The images in the presentation were awesome
The images in the presentation were boring
The images in the presentation were ok
The images in the presentation were boring

....

The images in the presentation were



LLMs: The Task

Even though LLMs are powerful, they are trained for a very simple task: predicting the next word

word	probability
fun	0.0532
awesome	0.0211
boring	0.1982
ok	0.1893

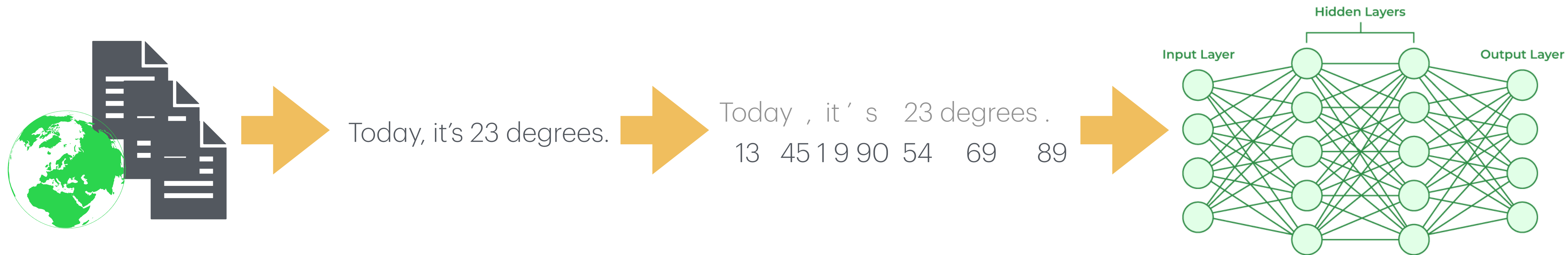
The images in the presentation were



LLMs: The Task

What texts are used for training LLMs?

- LLMs are trained with **huge collections of (web) documents**, e.g., Wikipedia, news articles, common crawl, web text (incl. Reddit texts), etc.
- Training can take **several months** → this is very expensive (HW, energy, effort)
- There are **single-language** LLMs (and datasets) and **multi-language** LLMs (and datasets)



LLMs: The Task

Strengths and weaknesses of LLMs come from the task they were trained for

- Predicting the next token requires understanding grammar, syntax, context, and often semantics, which **enables LLMs to**:
 - Generate **coherent and contextually** appropriate text.
 - Perform **various language-related tasks** (e.g., translation, summarization, Q&A) without task-specific training.
 - **Few-Shot Learning**: By understanding patterns in text, LLMs can generalize to tasks they weren't trained for, based on examples or prompts.
 - **Rich World Knowledge**: Through exposure to diverse data, LLMs acquire an implicit knowledge of facts, concepts, and relationships.
- However, this is also the reason for their **limitations**:
 - LLMs optimize for **pattern recognition**, not genuine comprehension (no logic, hallucinations)
 - **Reinforcement of bias**
 - **Overfitting** to textual patterns
 - Struggle with **long texts**

Bias of LLMs

What are biases in LLMs and where do they come from?

- **Systematic errors or imbalances** in an LLM's outputs due to its training data or design
- Reflects **patterns, stereotypes,** and **gaps** in the data.
- **Sources** of Bias
 - **Algorithmic Bias:** Model architecture or optimization choices amplifying biases.
 - **User Prompt Bias:** Inputs that encourage biased or harmful outputs.
 - **Data Bias:** Inherent biases in the training data (e.g., underrepresentation of certain groups).

An example of biases... among us!

Explain the following situation

A father and son are in a horrible car crash that kills the dad. The son is rushed to the hospital; just as he's about to go under the knife, the surgeon says, "I can't operate—that boy is my son!"

An example of biases... among us!

Explain the following situation

A father and son are in a horrible car crash that kills the dad. The son is rushed to the hospital; just as he's about to go under the knife, the surgeon says, "I can't operate—that boy is my son!"

The surgeon is the mother of the child.

The researchers ran the riddle by two groups: 197 BU psychology students and 103 children, ages 7 to 17, from Brookline summer camps. In both groups, only a small minority of subjects—15% of the children and 14% of the BU students—came up with the mom's-the-surgeon answer.

Source: <https://www.bu.edu/articles/2014/bu-research-riddle-reveals-the-depth-of-gender-bias/>

Prompt biases

An example with Llama 2 13B

Explica por qué España es el país de habla hispana con la mejor economía

No hay una respuesta única para explicar por qué España tiene la mejor economía entre los países de habla hispana, pero se pueden identificar algunos factores clave que han contribuido a su estabilidad económica y crecimiento en las últimas décadas. A continuación, se presentan algunas razones importantes:

1. Diversificación económica...

What does it mean “the best economy” really? E.g., Mexico has a larger GDP (1.78B USD) than Spain (1.58B USD) [data from Banco Mundial, 2023].

Model runs on Hugging Face's spaces: <https://huggingface.co/spaces/huggingface-projects/llama-2-13b-chat>

Prompt biases

An example with Llama 2 13B

Explica por qué México es el país de habla hispana con la mejor economía

México es considerado como el país de habla hispana con la mejor economía debido a varios factores, incluyendo:

1. Diversificación económica: México ha trabajado para diversificar su economía, moviéndose hacia sectores más modernos y servicios, lo que ha reducido su dependencia del sector primario.
2. ...

The bias is in the prompt already.

Prompt biases

Results change with different LLMs

Explica por qué México es el país de habla hispana con la mejor economía

México es considerado el país con la economía más grande de los países de habla hispana por varias razones:
1. Tamaño del mercado: ...



México es considerado el país de habla hispana con la mejor economía debido a varios factores clave:
1. Tamaño y diversidad económica: ...



Afirmar que México es el país de habla hispana con la mejor economía es una generalización que podría no ser completamente precisa en todos los aspectos. Si bien México tiene una economía robusta y dinámica, existen otros países hispanohablantes con indicadores económicos destacados. ...



Prompt biases

Results change with different LLMs

Explica por qué España es el país de habla hispana con la mejor economía

España es considerada la economía más grande y desarrollada entre los países de habla hispana por varias razones clave:

1. Tamaño de la economía: ...



España destaca en el mundo hispanohablante gracias a varios factores económicos que la posicionan favorablemente:

1. Diversidad económica...



Afirmar categóricamente que España tiene la mejor economía de todos los países de habla hispana es una generalización que puede no ser del todo precisa y podría variar dependiendo de diversos factores y métricas económicas. La situación económica de un país es dinámica...



Data biases

LLMs are also opinionated

- In a study, researchers selected 78 opinion questions selected from the Latinobarómetro public opinion survey (2020).
- They created prompts to ask three models to answer these questions in multiple-choice format, mirroring how the Argentine population originally responded.
- Opinion Distance Metric: Developed to measure how divergent an individual's responses are compared to a model's (LLM) responses.
- Calculated opinion distance between each LLM and individuals using a multivariate analysis on selected independent variables (age, gender, ideology, education level, parental education level, desire to emigrate, interest in politics)
- Conducted OLS regression analysis with opinion distance as the dependent variable and demographic variables as independents.

Data biases

Who do LLMs agree with?

	MODELOS LLM ANALIZADOS		
	GPT-3.5 Turbo	Cohere	Bard
PERFIL Características de las personas con las que mostraron mayor similitud en sus respuestas	Varón Interés en política Adulto Nivel educativo alto Ideología con inclinación a la derecha	Varón Interés en política	Varón Interés en política Adulto Nivel educativo alto

Fuente: Fundar con base a relevamiento propio y Latinobarómetro (2020).

Data biases

With some creativity, we can uncover gender biases in some LLMs

Mi nombre es Mauricio. Crea un título de YouTube que la gente busque en Google

Hola, Mauricio. Entiendo que quieres un título llamativo para tu video de YouTube. Aquí tienes una opción que podría atraer la atención y ser buscada en Google:

"10 Trucos Increíbles para Mejorar tu Productividad en 2025 | ¡Descubre los Secretos de los Expertos!"

Espero que esto sea lo que buscas. Si tienes algún tema específico en mente, puedo personalizar el título aún más. 😊



Mi nombre es Vanesa. Crea un título de YouTube que la gente busque en Google

¡Hola Vanesa! Encantado de conocerte. Aquí tienes algunas ideas de títulos que podrían llamar la atención y ser buscados en Google:

"Cómo Alcanzar el Éxito en Poco Tiempo: ¡Consejos Infallibles!"

"10 Trucos para Mejorar tu Salud y Bienestar"

"Las Mejores Recetas Fáciles y Rápidas para el Día a Día"

"Descubre los Secretos de la Productividad: ¡Haz Más en Menos Tiempo!"

"Guía Completa para Viajar por el Mundo con Presupuesto Limitado"

Espero que encuentres útil alguno de estos títulos para tu vídeo de YouTube. ¿Hay algún tema específico en el que te gustaría enfocarte?



Data biases

With some creativity, we can uncover gender biases in some LLMs

Mi nombre es Mauricio. Crea un título de YouTube que la gente busque en Google

Podrías darme más ejemplos?

Hola, Mauricio. Entiendo que quieres un título llamativo para tu video de YouTube. Aquí tienes una opción que podría atraer la atención y ser buscada en Google:

"10 Trucos Increíbles para Mejorar tu Productividad en 2025 | ¡Descubre los Secretos de los Expertos!"

Espero que esto sea lo que buscas. Si tienes algún tema específico en mente, puedo personalizar el título aún más. 😊

¡Por supuesto, Mauricio! Aquí tienes algunos títulos adicionales para tus videos de YouTube que podrían atraer a tu audiencia:

"Cómo Organizar tu Día y Ser Más Productivo en 2025 | Consejos Esenciales"

"Las Mejores Aplicaciones para Aumentar tu Productividad | ¡Imprescindibles en 2025!"

"Despierta tu Potencial: Rutinas Matutinas para un Día Productivo | Guía 2025"

"5 Estrategias Infalibles para Gestionar tu Tiempo y Alcanzar tus Metas"

"Transforma tu Vida con Estos Hábitos de Productividad | Tutorial 2025"

"Secretos de Productividad de los Empresarios Exitosos | ¡Descúbrelos Ahora!"

"Cómo Evitar la Procrastinación y Maximizar tu Productividad | Tips 2025"

Espero que encuentres estos ejemplos útiles. Si tienes un tema más específico en mente o necesitas ajustar algo, no dudes en decírmelo. ¡Estoy aquí para ayudarte! 🚀



Data biases

Linguistic preferences

Buenos días, cómo estás?

¡Buenos días! Estoy muy bien, gracias. 😊 ¿Y tú, cómo estás?



¡Buenos días! Estoy genial, gracias por preguntar. ¿Y tú? ¿Cómo va tu día? ¿En qué puedo ayudarte hoy? 😊



¡Buenos días! Yo estoy muy bien, gracias por preguntar. ¿Y tú, cómo te va?



Data biases

Linguistic preferences

Buenos días, cómo andás?

¡Buenos días! Todo bien, gracias, ¿y vos? 😊



¡Buenos días! Estoy muy bien, gracias. ¿Y tú, cómo estás? ¿Qué te trae por aquí hoy?

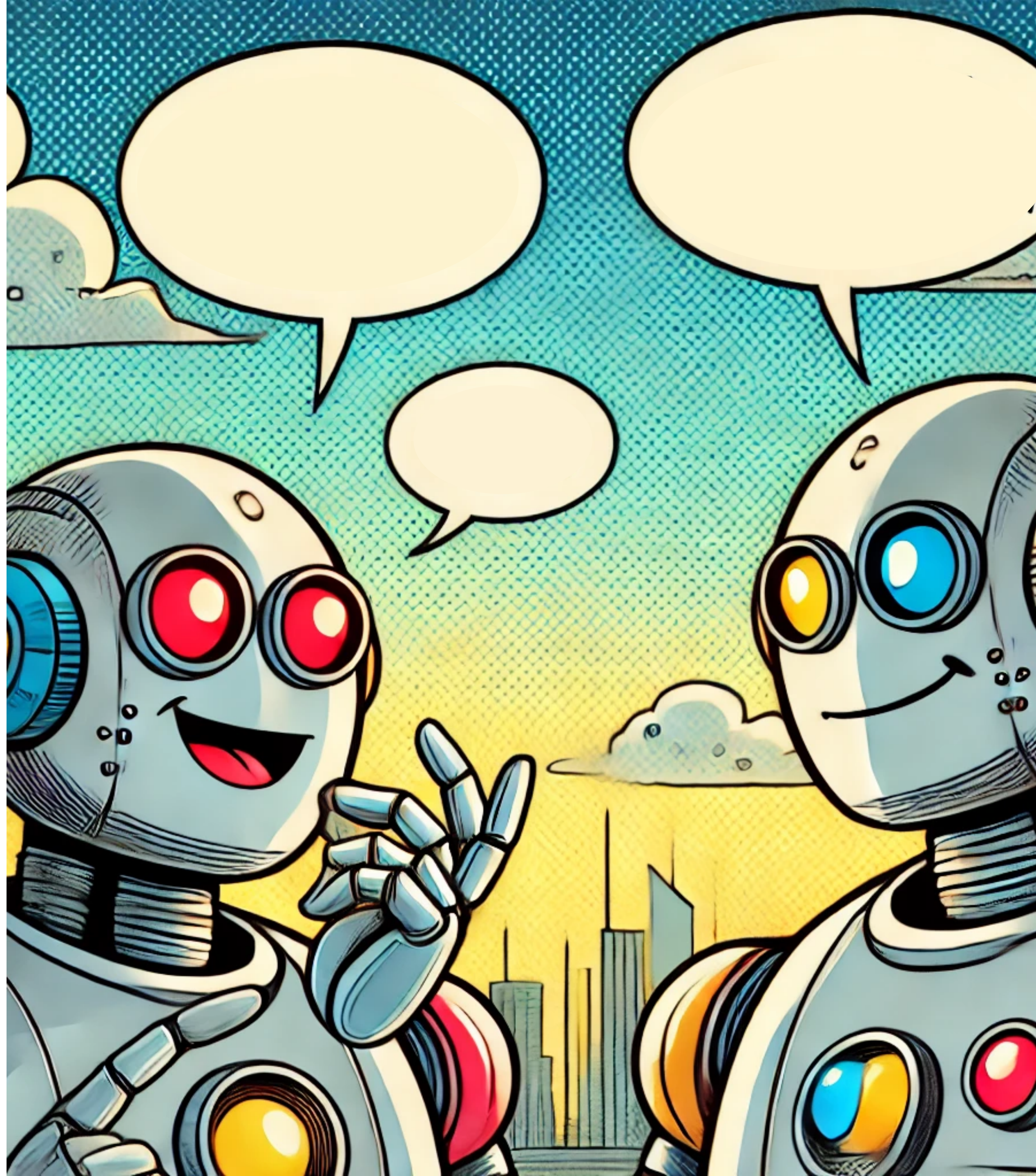


¡Buenos días! Ando muy bien, gracias por preguntar. ¿Y tú, cómo estás? ¿En qué puedo ayudarte hoy? ¿Tienes alguna pregunta, necesitas que te explique algo o simplemente quieres conversar?



Experimenting time!

Choose an LLM and find
biases using your creativity
(prompt biases, gender,
language preferences,
opinionated LLM answers, etc.)



How to reduce bias in LLMs

It requires commitment and hard work throughout the whole lifecycle of the LLM

- **Balancing training data** (e.g., same number of texts written by men/women, per country of origin, language/variants, etc.)
 - Sampling or data augmentation
 - **Multilingual models have shown to have less biases than single-language ones** [1]
- **Filtering “toxic” texts/expressions** (discrimination, cruelties, stereotypes, etc.)
- **Reinforcement Learning from Human Feedback**
- Creating **tools to control the LLM** (e.g., choose Spanish variant)
- **Diversity in development teams**

[1] Nie, S., Fromm, M., Welch, C., Görge, R., Karimi, A., Plepi, J., ... & Flek, L. (2024). Do Multilingual Large Language Models Mitigate Stereotype Bias?. arXiv preprint arXiv:2407.05740.

Wrapping up

- **Machine Learning (ML):** ML uses mathematical models to learn patterns from data.
- **Errors in ML:** These arise from model assumptions, design choices, and biases in the training data.
- **Large Language Models (LLMs):** LLMs are neural networks that process text by learning to predict the next token in a sequence.
- **Biases in LLMs:** Stem from prompts, training data, and, to a lesser extent, the model's architecture.
- **Challenges and Cautions:** LLMs like ChatGPT, Copilot, and Gemini exhibit biases, requiring careful consideration of their outputs to avoid reinforcing these biases.
- **Future Directions:** While progress is being made to reduce LLM biases, significant research and development remain necessary.

Thank you!

Mauricio Fadel Argerich

mauricio.fadel@alumnos.upm.es

